LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-PROC-652239

Image Segmentation using Consensus from Hierarchical Segmentation Ensembles

H. Kim, J. J. Thiagarajan, P. Bremer

March 25, 2014

IEEE International Conference on Image Processing
Paris, France
October 27, 2014 through October 30, 2014

Disclaimer

IMAGE SEGMENTATION USING CONSENSUS FROM HIERARCHICAL SEGMENTATION ENSEMBLES

Hyojin Kim, Jayaraman J. Thiagarajan and Peer-Timo Bremer
Lawrence Livermore National Laboratory, Livermore, California, USA

ABSTRACT

Unsupervised, automatic image segmentation without contextual knowledge, or user intervention is a challenging problem. The key to robust segmentation is an appropriate selection of local features and metrics. However, a single aggregation of the local features using a greedy merging order often results in incorrect segmentation. This paper presents an unsupervised approach, which uses the consensus inferred from hierarchical segmentation ensembles, for partitioning images into foreground and background regions. By exploring an expanded set of possible aggregations of the local features, the proposed method generates meaningful segmentations that are not often revealed when only the optimal hierarchy is considered. A graph cuts-based approach is employed to combine the consensus along with a foreground-background model estimate, obtained using the ensemble, for effective segmentation. Experiments with a standard dataset show promising results when compared to several existing methods including the state-of-the-art *weak* supervised techniques that use co-segmentation.

***Index Terms*—** Unsupervised segmentation, multiple hierarchies, consensus clustering, graph cuts, superpixels.

1. INTRODUCTION

Segmenting images into foreground and background regions is a long-standing and difficult problem in computer vision. Since foreground regions in natural images often contain objects and well-defined structures, it is natural to employ supervised methods for segmentation [1, 2, 3]. However, this requires a large collection of training images with manual segmentations, and it might be expensive and difficult to obtain enough samples. As a result, semi-supervised methods, that require some level of user intervention, have gained popularity [4, 5]. Though this relaxes the requirement of a large training set, user intervention is required for every image and this dependency makes such methods ineffective in several cases. Recent approaches [6, 7, 8, 9] build a *weak* supervised set for each image, by selecting visually similar images from an external database, and perform co-segmentation of all images together. The effectiveness of these methods depends on the intra-class heterogeneity of the *weak* supervised set.

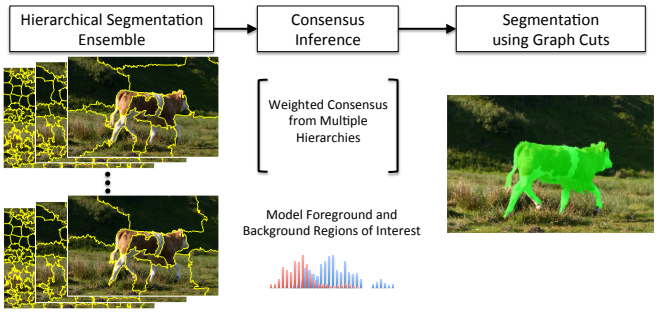

Fig. 1. An overview of the proposed algorithm for unsupervised segmentation.

In this paper, we focus on the problem of unsupervised, single image segmentation. Though unsupervised segmentation approaches are attractive, they depend strongly on the local features in an image, and often cannot accurately identify larger coherent regions. Furthermore, the need for an appropriate choice of feature/metric and parameters makes the design of automatic segmentation methods very challenging. An important class of unsupervised segmentation methods employs graph partitioning to identify foreground and background regions in an image. In [10], the authors developed the multiscale normalized cut algorithm that incorporates a multiscale graph structure to regularize the graph partitioning problem to improve single image segmentation. Furthermore, this approach was extended to the case of hierarchical segmentation in [8], where parent-child relationships between segments from different levels of the hierarchy were used for regularization. Since no effective methods exist for choosing optimal parameters in unsupervised segmentation, there is also a growing interest to combine several segmentations obtained using different parameter settings or different segmentation algorithms into a final consensus segmentation [11, 12, 13]. The use of ensemble methods in unsupervised learning has enabled robust estimation of the number of clusters and the underlying partitioning [14].

In the proposed work, we investigate the use of hierarchical segmentation ensembles, which randomizes the order of merging regions while building a hierarchy. Based on a consensus inferred from the multiple hierarchies, we propose to perform image segmentation using a graph cuts approach. We

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-PROC-652239.

evaluate the performance of the proposed algorithm using the MSRC objects dataset [15], and show that our method provides promising results, in comparison to existing algorithms.

2. AN OVERVIEW OF THE PROPOSED APPROACH

Figure 1 illustrates an overview of our approach for automatic partitioning of foreground and background regions. As it is common with several state-of-the-art methods, our algorithm begins with a set of superpixels [16] (perceptually meaningful atomic regions, well-aligned with edges) extracted from an input image, and builds an ensemble of hierarchical segmentations. The multiple hierarchies explore a number of possible aggregations by randomly shuffling the set of candidate region pairs that can be merged, within a threshold. The hierarchical ensemble is then aggregated by constructing a weighted consensus matrix. Finally, a graph cuts-based procedure, that exploits both the consensus information and a foreground-background model estimated using the top-level segments, is adopted to partition the image.

The novelty of our algorithm lies in applying multiple hierarchies to the consensus voting. The multiple hierarchies overcome the issue of allowing only a single aggregation at each level. Contrary to existing consensus clustering schemes that combine segmentations obtained using different parameter settings or algorithms, we focus on order dependency, which is a critical issue in hierarchical segmentation algorithms. Some candidate region pairs might end up being unmerged, due to the arbitrary order in which we merge the local regions. Using multiple hierarchies resolves this issue by randomizing the merging order to explore as many aggregations as possible. Our method also takes into account the hierarchical relationships between regions during consensus inference.

Furthermore, similar to the successful semi-supervised approaches, our method also incorporates a foreground-background model estimate into the graph cuts formulation. Though the model estimation can be performed using the superpixels directly, we show that the ensemble results give a better understanding of the underlying contextual structure, resulting in a more robust segmentation.

3. HIERARCHICAL SEGMENTATION ENSEMBLE

The first step in the proposed algorithm generates an ensemble of hierarchical segmentations from an initial set of superpixels. We denote the superpixel set by $V_0 = \{x_i\}_{i=0}^{N_0-1}$ where N_0 is the total number of superpixels. The set of regions in any level of a hierarchy is accompanied by a Region Adjacency Graph (RAG) describing the spatial adjacency among the regions. We denote the undirected RAG for the set of superpixels by $G_0 = (V_0, E_0)$ where E_0 is the set of edges for the nodes in V_0. Each hierarchical segmentation incrementally constructs RAGs while merging regions from the previous level (initially superpixels). The resulting hierarchy can be represented as a collection of RAGs, $\{G_\ell\}_{\ell=0}^{L-1}$, where L is the number of hierarchy levels, and $G_\ell = (V_\ell, E_\ell)$, for $\ell = 0 \cdots L-1$. Denoting the number of regions in level ℓ by N_ℓ, the region $r_i^\ell \in V_\ell$ corresponds to ith region at that level, and $e_{i,j}^\ell \in E_\ell$ is an adjacency edge connecting r_i and r_j, with the corresponding edge weight (similarity) $w_{i,j}^\ell$.

In each level of a hierarchy, the edge weights between regions are computed based on three different terms as

$$w_{i,j}^\ell = \prod_{n=1}^{3} W_n(i,j,\ell). \tag{1}$$

The first term measures the similarity between the color histograms of two regions as

$$W_1(i,j,\ell) = \min_{x_n \in r_i^\ell, x_m \in r_j^\ell} \exp\big(-\sigma_1 \chi^2(H(x_n), H(x_m))\big),$$

where $H(x_n)$ is the 3-channel LAB color histogram (64 bins for each channel) of superpixel x_n and χ^2 measures the chi-square distance between two histograms. The second term for obtaining the ratio of edge pixels is defined as

$$W_2(i,j,\ell) = \exp\left(-\sigma_2 \frac{G(i,j,\ell)}{B(i,j,\ell)}\right),$$

where $G(i,j,\ell)$ measures the number of border pixels on or near an Canny edge, and $B(i,j,\ell)$ is the total number of pixels in the border shared between r_i^ℓ and r_j^ℓ. The third term is obtained by normalizing the number of border pixels shared between two regions by the total number of border pixels as

$$W_3(i,j,l) = \exp\left(-\sigma_3 \left(1 - \max\left(\frac{B(i,j,l)}{B_t(i,l)}, \frac{B(i,j,l)}{B_t(j,l)}\right)\right)\right),$$

where $B_t(i,l)$ returns the total number of border pixels for the region r_i^ℓ. Here, σ_1, σ_2, and σ_3 are parameters of the Gaussian radial basis functions.

Two spatially adjacent regions r_i^ℓ and r_j^ℓ in a level are considered homogenous, and consequently $e_{i,j}^\ell$ becomes a candidate edge if $w_{i,j}^\ell \geq \delta$ where δ is a predefined threshold. The set E_ℓ^C contains the candidate edges at level ℓ sorted in the descending order based on their edge weights. Two homogenous regions, r_i^ℓ and r_j^ℓ, are merged if $e_{i,j}^\ell \in E_\ell^C$, and either of them has not been merged previously.

In our setup, we propose to explore all potential aggregations by randomizing the merging order of the candidate edges to find an optimal segmentation. Extending our notation, we define the collection of RAGs for the multiple segmentations as $\{\{G_\ell^m\}_{\ell=0}^{L-1}\}_{m=0}^{M-1}$, where M is the number of segmentations in the ensemble. We denote the number of levels in the hierarchical segmentation m by L_m. In all cases, the merge process to build a hierarchy continues until there is no homogenous region pair to be merged or a stopping criterion on the number of levels is met.

(a) (b) (c)

Fig. 2. Segmentation results obtained with different formulations for graph cuts. (a) N-cut [17] using the consensus to define only the smoothness term, (b) Combining the consensus information with a foreground-background model estimated from the superpixels directly, (c) Proposed method uses the ensemble to estimate the model, in addition to the consensus.

4. SEGMENTATION USING GRAPH CUTS

Each hierarchy in the ensemble provides a meaningful partitioning of the image, and we propose to build a consensus, using the ensemble, which represents the probabilities of merging any superpixel pair. We adopt a strategy similar to graph cuts, where the consensus information is combined with an estimate of the foreground-background model, to determine the optimal partitioning. Though this algorithm can be easily generalized to work for any number of segments, for simplicity, we consider the case of $K = 2$, i.e., separating a foreground from its background.

4.1. Consensus Inference

We perform consensus inference from an ensemble of M hierarchical segmentations of the same image. Unlike conventional consensus clustering, we need to take into account the hierarchical relationships. We first initialize the consensus matrix $\mathbf{C} \in \mathbb{R}^{N \times N}$ to zeros. The coherence between the superpixels x_i and x_j can be estimated from the ensemble as

$$c_{ij} = \frac{1}{M} \sum_{m=0}^{M-1} \sum_{\ell=0}^{L_m - 1} \frac{\Gamma(G_\ell^m, i, j)}{L_m}, \tag{2}$$

where

$$\Gamma(G_l^m, i, j) = \begin{cases} 1, & \text{if } \exists r \in G_\ell^m, x_i \subseteq r \wedge x_j \subseteq r, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

We propose to use the consensus matrix to impose the smoothness penalty in the graph cuts formulation.

4.2. Foreground-Background Segregation

This section describes an algorithm to perform foreground-background segmentation, given the hierarchical segmentation ensemble and the consensus matrix. Typically, graph cuts-based algorithms are semi-supervised, where user intervention (drawing a bounding box or selecting regions of interest) is required to estimate models for background and the foreground. We reformulate this approach by incorporating information from the ensemble of segmentation hierarchies. Following [4], we define an Markov Random Field (MRF) based cost function for graph cuts as

$$F(\mathbf{A}) = \sum_{x_i \in V_0} F_d(\alpha_i) + \lambda \sum_{x_i, x_j \in V_0} F_s(\alpha_i, \alpha_j), \tag{4}$$

where $\mathbf{A} = [\alpha_0, \cdots, \alpha_{N_0 - 1}]^T$ is a vector of binary labels (foreground and background) and α_i denotes the label for superpixel $x_i \in V_0$. Note that each node in the graph is a superpixel and the corresponding edge weights are defined using the consensus matrix $\mathbf{C}$. The data cost, F_d, defines the penalty to assign a label to each superpixel, and F_s is the smoothness cost that penalizes a pair of labels assigned to connected superpixels. When only the smoothness cost is considered in (4), it is equivalent to performing normalized cut with the consensus matrix. Figure 2(a) shows an example partitioning, that uses only the consensus information, and we can clearly see that the segmentation is suboptimal.

Though the superpixels can be directly used to estimate the statistics of foreground and background regions in the im-

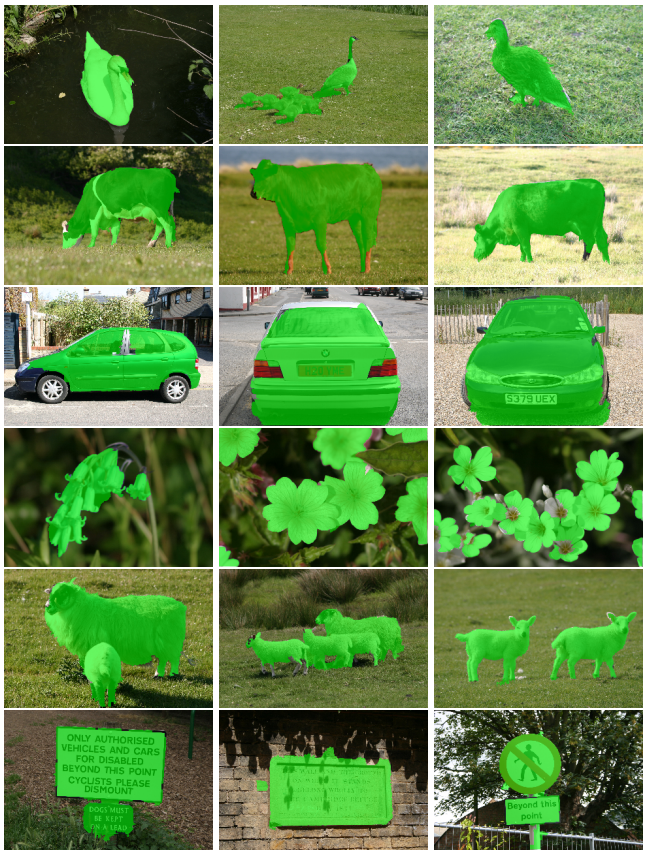

Fig. 3. Segmentation results of the proposed algorithm for an example set of images from the MSRC dataset.

Table 1. Performance on the MSRC dataset (*Jaccard Coefficient*), compared to other state-of-the-art approaches.

	MNCut [10]	**HNCut** [8]	**CoSand** [7]	**HCoSeg** [8]	**Ours**
Bike	40.8	39.5	42.3	42.1	**43.9**
Bird	28.1	29.5	31.7	32.8	**56.7**
Car	43.5	49.5	**56.2**	54.4	55.6
Cat	37.6	40.3	41.7	44.6	**48.3**
Chair	33.2	41	39.9	42.9	**52.8**
Cow	38.9	50.8	40.1	52.3	**67.2**
Dog	32.2	38.9	41.9	42.1	**54.1**
Face	33.9	35.5	36.7	37.6	**51.0**
Flower	45.1	53.7	53.8	58.9	**68.6**
Plane	27.3	29.5	**35.1**	32.7	30.2
Sheep	41.7	59.1	43.8	62.1	**75.6**
Sign	58.8	60.1	51.7	53.3	**67.9**
Tree	47.3	58.5	58.9	**61.2**	59.8

age, when completely unsupervised, the model estimation is not robust (Figure 2(b)). Since, contextual knowledge of the foreground regions is needed to define a robust data cost, we propose to use the regions from the top level of all hierarchies in the ensemble. We consider both the color histogram (64 bins for each channel) of the region, and the normalized distance between the region center and the center of the image (for reference) to understand the foreground/background separation. Following this, we categorize the regions into $K = 2$ groups using unsupervised clustering, where one cluster describes the foreground and the other the background. The location information helps to achieve spatial coherence. Although one can employ sophisticated modeling techniques, we use the simple K-means clustering method. Given the two cluster centroids H_0 and H_1, the data cost $F_d(\alpha_i)$ is measured as $\exp\big(-\gamma\chi^2(H(x_i), H_0)\big)$, and $\exp\big(-\gamma\chi^2(H(x_i), H_1)\big)$ for the foreground and background labels respectively. The improvement in segmentation achieved by using the ensemble to build the data cost is evidenced in Figure 2(c).

5. RESULTS AND DISCUSSION

We performed experiments on the MSRC dataset [15], containing 20 classes of images with ground truth segmentations. We extracted SLIC Superpixels [16] from every image to initialize our segmentation algorithm. To avoid crossing over strong region boundaries, over segmentation was performed ($N_0 \approx 200$, compactness = 10). In all cases, the penalty λ in (4) was fixed at 0.5. For quantitative evaluation, we selected 13 classes out of the dataset, and measured the segmentation accuracy with respect to the ground truth labels. The Jaccard coefficient between two regions, computed as $(R_A \cap R_B)/(R_A \cup R_B)$, measures the ratio of overlap.

Table 1 summarizes the average overlap scores obtained for each class, with two unsupervised segmentation methods [10, 8] and two *weak* supervised co-segmentation methods [7, 8]. We fixed $M = 50$, $L = [10..20]$. The results in Table 1 demonstrate the superior performance of our method in most cases, even when compared to co-segmentation methods. With several images containing relatively complex textures and content, our method successfully extracted the foreground regions, as illustrated in Fig. 3.

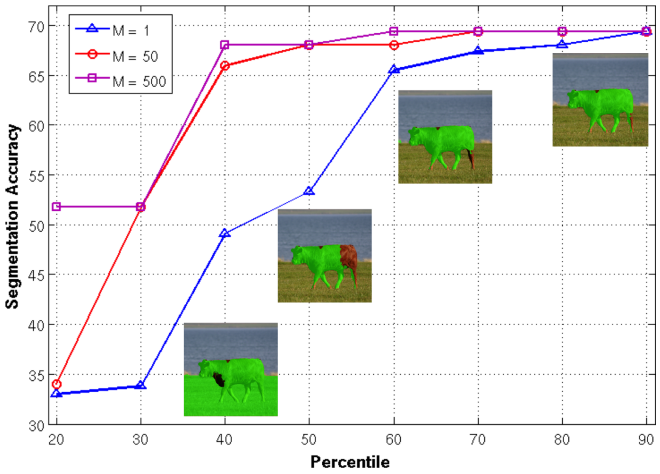

Fig. 4. Segmentation accuracy measured against the number of hierarchies, M, for an example image. In each case, we repeated the algorithm for 100 trials and report the performance at different percentiles to understand the worst-case behavior.

The choice of the number of hierarchies in the ensemble, M, determines the computational complexity of the algorithm. We used an example image from the MSRC dataset to demonstrate the dependency of the segmentation accuracy on M (Fig. 4). We consider the cases $M = \{1, 50, 500\}$ and since the proposed approach is randomized, we repeat the algorithm for 100 trials. Though it is common to look at the average behavior, we believe it is crucial to understand the worst-case behavior for the different parameter settings. We observe that, as M increases, the worst-case behavior improves significantly. We compute the segmentation accuracies at different percentiles, i.e. performance at the r^{th} percentile implies that it is better than proportion r of the trials. At the 90^{th} percentile, the algorithm provides high quality segmentation in all cases, whereas the median (50^{th} percentile) performance for $M = 1$ drops significantly when compared to the other cases. In particular, the worst case segmentation obtained with $M = 500$ is as good as the median performance with $M = 1$. Performance of the segmentation with $M = 50$ matches that of $M = 500$ except at very low quantiles.

In summary, building an ensemble of hierarchical segmentations, by ignoring the order-dependency of the merging, enables us to perform meaningful partitioning of images. The consensus inference holds the key to the performance of the approach and more sophisticated strategies can be developed. The initial results are very promising, and our approach performs better than methods that perform co-segmentation of visually relevant images from a large database.

6. REFERENCES

[1] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proc. of IEEE CVPR*, June 2008, pp. 1–8.

[2] J.M. Gonfaus, X. Boix, J. van de Weijer, A.D. Bagdanov, J. Serrat, and J. Gonzalez, "Harmony potentials for joint classification and segmentation," in *Proc. of IEEE CVPR*, June 2010, pp. 3280–3287.

[3] Aurelien Lucchi, Yunpeng Li, Xavier Boix Bosch, Kevin Smith, and Pascal Fua, "Are spatial and global constraints really necessary for segmentation?," in *Proc. of IEEE ICCV*, 2011, pp. 9–16.

[4] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images," in *Proc. of IEEE ICCV*, 2001, pp. 105–112.

[5] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics (SIGGRAPH'04)*, vol. 23, pp. 309–314, 2004.

[6] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Proc. of IEEE CVPR*, 2010, pp. 1943–1950.

[7] G. Kim, E. Xing, L. Fei-Fei, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *Proc. of IEEE ICCV*, 2011, pp. 169–176.

[8] E. Kim, H. Li, and X. Huang, "A hierarchical image clustering cosegmentation framework," in *Proc. of IEEE CVPR*, 2012, pp. 686–693.

[9] Fan Wang, Qixing Huang, and Leonidas J. Guibas, "Image co-segmentation via consistent functional maps," in *Proc. of IEEE ICCV*, 2013.

[10] T. Cour, F. Benezit, and J. Shi, "Spectral segmentation with multiscale graph decomposition," in *Proc. of IEEE CVPR*, 2005, pp. 1124–1131.

[11] Shih-Hung Chen, Ming-Jui Kuo, and Jung-Hua Wang, "Image segmentation based on consensus voting," in *International Workshop on Cellular Neural Networks and Their Application*, May 2005, pp. 1–4.

[12] Lucas Franek, DanielDuarte Abdala, Sandro Vega-Pons, and Xiaoyi Jiang, "Image segmentation fusion using general ensemble clustering methods," in *Proc. of ACCV*, 2011, vol. 6495, pp. 373–384.

[13] Mete Ozay, Sanjeev R. Kulkarni, and H. Vincent Poor, "Context-based text detection in natural scenes," in *Proc. of IEEE ICIP*, 2012, pp. 1857–1860.

[14] Carl Dean Meyer, Shaina Race, and Kevin Valakuzhy, "Determining the number of clusters via iterative consensus clustering," in *SDM*. 2013, pp. 94–102, SIAM.

[15] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. of IEEE CVPR*, 2005, pp. 1800–1807.

[16] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 2274–2282, 2012.

[17] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 888–905, 2000.